Appendix 3. Data Engineering Considerations

Data Engineering	
Step	**Purpose**
Identification	Identify each data element that is to be moved into the DSS, as well as the data source for this information.
Primacy	If multiple data elements are identical by name, content, or alias, select one as the primary element. If conflicts arise between the contents of these elements, one will automatically be deemed correct.
Resolution	When a data element name is used by more than one data source and the content of data elements is fundamentally different, map one or both of the data elements to new names. Fundamentally, "different" means that the content of the elements refers to different concepts. For example: Table1.locale = North (a geographic direction) Table2.locale = Region IV (an office)
Structure	When the same data elements (representing the same concept) from different sources are stored in different formats, select one format as your standard, and convert the others to it. For example, a date may be stored in one source as a text string ("Wednesday, May 8, 2002"), while it may be stored in another source as the computed value ("37384").
Audit	If possible, audit a formatted report of incoming data. (This may involve only a sample of a large data dump.) An audit will tell you if checks on the source system are working improperly or if they are being circumvented. For example, you may find 00000 or 99999 in the ZIP code field or 800-555-1212 as a phone number. An audit does not check whether an individual record is correct.
Point of Contact	Associate each data source with contact information. Request this information in every data source contract, Memorandum of Understanding (MOU), or Memorandum of Agreement (MOA).
Error Reporting	Include a mechanism in each MOA or contract for passing error information back to the source authority. Errors refer to things such as eight-digit Social Security numbers, four-digit ZIP codes, and NCPs who are 6 years old. An error-reporting mechanism tells you: ■ How the information will be passed back to the source authority ■ What initial conditions will cause a record to be rejected and logged during Stage 1 cleansing ■ How initial records with nonfatal errors will be processed during the Extract, Transform, and Load (ETL) process. The word "initial" refers to the SCS-DSS design period. When live data begins to flow through the system, the error-reporting plan will develop many more conditions.
Size/Population	Estimate the size of your system's initial storage requirements for performing Stage 1 cleansing, completing ETL, loading the Operational Data Store (ODS), and populating the data mart. Calculate the population by estimating the number of records stored in the ODS and data mart.
ETL	Determine the time allowed to complete the ETL process. In one particularly bad implementation of a DSS, ETL required 3 months for a single month of data. ETL encompasses: ■ *Extraction time*—The time required to move the data from its source to the DSS. Of the three ETL processes, this one usually requires the most time. ■ *Transformation time*—This includes Stage 1 cleansing, transforming the data into the appropriate architecture for your SCS-DSS, and loading it into the ODS. ■ *Load time*—The time required to copy data from the ODS into the schemas within the data mart. There are three ways to adjust these processes to meet ETL deadlines: ■ *Increase your processing power*—This may not be a viable solution if the bottleneck is outside your control. For example, source owners may be unwilling or unable to upgrade their system capacities. ■ *Stagger your schedule*—If a weekly update of your SCS-DSS is acceptable, you may be able to transfer data in smaller chunks throughout the week. This reduces the processing load at any one time. The only potential problem is that source systems may not be able to change their schedules to meet your mini-deadlines. ■ *Use intelligent design*—There are many types of databases; make sure your design is appropriate for the type of database you want and for the manner in which you intend to use it.